The relative responsiveness of test instruments can be estimated using a meta-analytic approach: an illustration with treatments for depression.

D. Kounali[1], K. Button[1,3], G. Lewis[2] , A. E. Ades[1]

[1]School of Social and Community Medicine, University of Bristol, 39 Whatley Road, Bristol BS8 2PS, United Kingdom

[2]School of Life and Medical Sciences, Faculty of Brain Sciences, Division of Psychiatry, University College London, 67-73 Riding House St., London W1W 7EJ, UK

[3]Department of Psychology, 6.19 Wessex House, University of Bath, Bath BA2 7AY, UK

Abstract

Objective. We present a meta-analytic method that combines information on treatment effects from different instruments from a network of randomized trials to estimate instrument relative responsiveness.

Study design. Five depression-test instruments (Beck Depression Inventory (BDI I/II), Patient Health Questionnaire (PHQ9), Hamilton Rating for Depression (HAMD 17 and 24-item), Montgomery-Asberg Depression Rating (MADRS)), and three generic quality of life measures (EuroQoL (EQ-5D); SF36 Mental (SF36-MCS) and Physical Component Summaries (SF36-PCS)) were compared. Randomized trials of treatments for depression reporting outcomes on any two or more of these instruments were identified. Information on the within-trial ratios of standardized treatment effects was pooled across the studies to estimate relative responsiveness.

Results. The between-instrument ratios of standardized treatment effects vary across trials, with a coefficient of variation of 13% (95% CrI 6%-25%). There were important differences between the depression measures, with PHQ9 being the most responsive instrument, and BDI the least. Responsiveness of the EQ-5D and SF36-PCS were poor. **SF36-MCS performed similarly** to depression instruments.

Conclusion. Information on relative responsiveness of several test instruments can be pooled across networks of trials reporting at least two outcomes, allowing comparison and ranking of test instruments that may never have been compared directly.

Keywords: Relative Responsiveness, test instruments, meta-analysis, depression**.**

1. Introduction

The responsiveness of a measuring instrument is broadly understood as its ability to detect meaningful changes in a patient's clinical state. However, there is little consensus on the mathematical way to determine the magnitude of change and a variety of indices have been used.

Comparative studies of the responsiveness of patient- and clinician-reported outcomes are undertaken to improve the assessment of interventions and inform the choice of outcome measure for future studies. The responsiveness of depression-specific instruments reported by either patients or clinicians, have been examined mainly in isolated studies. A variety of definitions exist [1, 2]. Some of them are based on assessing longitudinal within-patient changes in clinical condition over time, others on the ability to discriminate cross-sectionally between groups of patients, including patients in different arms in randomized trials. Lack of comparability in metrics poses challenges for evaluating interventions, comparing effects across different studies and designing new studies.

In this paper we focus on responsiveness to treatment in the context of clinical trials, but the same methods can be extended to longitudinal measures (see discussion). Measures of the *relative* responsiveness to the changes caused by interventions are of particular significance for the evaluation of therapeutic interventions. More importantly, relative instrument responsiveness measurers can **help guide** the choice of outcome measures in clinical trials. More responsive outcomes permit smaller trials at a given power, improving efficiency and reducing costs.

Whether longitudinal or cross-sectional designs are used, responsiveness is measured predominantly by "effect size" statistics, which are standardized by dividing the mean longitudinal change or a mean treatment difference by an estimate of the relevant standard

deviation (SD) [3-5]. The resulting standardized estimates, also known as "standardized effect sizes" are among the recommended criteria for identifying and selecting instruments by the Patient reported Outcomes Measurement Group [6]. The two main statistical standardization approaches are the Cohen's *d* [7] which involves division by a pooled SD, and Hedge's *g* [8] which includes a bias correction due to small sample size that is important for sample sizes less than 10 [9].

Significance of change scores are typically computed from paired t-tests. The ratio of effect sizes from two instruments is the ratio of paired t-tests. The ratio of the effect sizes from two test instruments can be interpreted as a measure of relative responsiveness. It estimates the extent to which one scale is more or less efficient at detecting change over time relative to another scale [10] .

Papers exploring responsiveness of one or more measures can be found in every field of medicine [11]. However, with some exceptions [12-14], the literature on comparative responsiveness of test instruments does not attempt to pool estimates across studies. Here we develop a formal meta-analytic approach, which pools within-trial information on the ratios of standardized effect sizes, across a connected network of trials. This allows investigators to draw conclusions about the relative responsiveness of several outcomes measures, including those that may never have been compared directly, and to rank test instruments in terms of responsiveness.

We begin by describing an illustrative dataset of trials of treatments for depression, in which outcomes were reported on five disease-specific and three generic Quality of Life (QoL) measures. We then explain the statistical methods, which have been used previously for simultaneously estimating treatment effects and "mappings" between outcomes on different scales [15, 16]. These methods derive from a common factor theory of test instruments [17, 18]. In the discussion section we describe the properties of our method, the assumptions being made, and its limitations. We also consider the proposed methods in relation to previous methodological work on responsiveness.

2. METHODS

2. 1 Illustrative dataset

A generic search for depression outcomes was initially conducted in May 2011 including all measures of depression, anxiety and quality of life available in studies on the Cochrane Depression Anxiety and Neurosis (CCDAN) Review Group's register. Details of CCDAN's generic search strategies can be found on the Group's website[1]. This search identified 75 studies reporting statistics with at least two of the following eight test instruments, five disease-specific and three health-related Quality of Life (QoL) scales: the Beck Depression Inventory (BDI I/II) [19], Patient Health Questionnaire (PHQ9) [20], Hamilton Rating for Depression scale (HAMD 17 and 24-item) [21], Montgomery-Asberg Depression Rating Scale (MADRS) [22]; EQ-5D or EuroQol [23]; and the "short-form" SF36 Mental and Physical Components Summaries [24]. Among these we identified 31 placebo- and/or usual care controlled studies with clearly defined treatment and control groups that reported *either* statistics at follow-up for each of comparison groups (i.e. means, SDs and sample sizes), *or* change score statistics (mean, SD, and sample size). Where both were reported, the follow-up scores were used in the analysis.

One of the 31 studies reported outcomes separately for patients at two levels of depression. We treated this as two independent trials, effectively giving 32 studies. Eleven were drug trials, and the remaining were studies of psychological therapies (psychotherapy, cognitive behavioral therapy, psychoeducation and problem solving in primary care settings). There were also four studies with assorted treatments including regimens involving monitoring or choice of pharmacological treatment, herbal medicine, and other treatments. A detailed description and citation listing can be found in the Supplementary materials (A.1). The treatment effects are expressed as differences relative to the control arm (Table 1). It should be noted that the criteria for study inclusion is far more liberal than would be appropriate for studies of relative treatment effects, as we will be pooling information on effect size ratios from a very wide range of treatments (see discussion).

[1] http://cmd.cochrane.org/search-strategies-identification-studies

Table 1: Treatment effects, their standard errors and pooled SD at follow-up, in the 32 included Studies for treatment of depression. SE Standard Error of mean treatment effect, SD standard deviation at follow-up or change score. (N control and N Treat numbers of patients in control and treatment arms, N number of observations for each outcome (including different arms contributing data on the same outcome).

Study Reference (Total N=111) (N control / N Treat.)	BDI Mean (SE) SD (pooled) N=25	PHQ9 Mean (SE) SD (pooled) N=9	HAMD17 Mean (SE) SD (pooled) N=35	HAMD24 Mean (SE) SD (pooled) N=2	MADRS Mean (SE) SD (pooled) N=21	EQ5D Mean (SE) SD (pooled) N=4	SF36-Mental Mean (SE) SD (pooled) N=9	SF36-Physical Mean (SE) SD (pooled) N=6
1. AIM 2002		-3.40 (0.84)	-3.30 (1.09)					
(93/89)		SD=5.70	SD=7.37					
2. Carney 2006	1.30 (1.80)		0.60 (1.20)					
(60/62)	SD=9.96		SD=6.59					
3. Ijff 2007	-2.80 (2.50)		-4.50 (2.24)					
(20/20)	SD=8.95		SD=8.67					
4. Konig 2009	1.87 (1.20)					-0.02 (0.07)		
(150/156)	SD=10.46					SD=0.65		
5.MIND IT (2002 b	-3.78 (2.13)		-2.44 (1.26)					
(41/37)	SD=8.07		SD=6.27					
6. Moak 2003	-2.10 (2.19)		-1.00 (1.48)					
(44/38)	SD=10.12		SD=6.63					
7. Ooskooilar 2006			-1.80 (0.85)		-3.30 (1.13)			
(199/198)			SD=8.45		SD=11.27			
8. Pope 2010			-0.60 (1.58)		-0.50 (2.05)			
(34/40)			SD=6.91		SD=8.89			
9. SADHART 2002	-0.70 (0.85)		-0.80 (0.57)					
(183/184)	SD=8.13		SD=5.42					
10. de-Battista 2010		-4.30 (1.74)			-5.40 (2.98)			
(44/34)		SD=7.76			SD=13.30			
11. ARISE-RD 2003			-5.70 (0.46)		-8.20 (0.63)			
(489/383)			SD=6.62		SD=9.11			
12. FAVA 2005			-1.20 (0.35)		-1.40 (0.77)			
(149/151)			SD=4.32		SD=8.67			
13. Mannel 2010		-1.40 (0.59)	-1.10 (0.61)					
(100/100)		SD=4.20	SD=4.30					
14. IPCRESS 2009	-7.50 (1.73)					0.07 (0.03)		
(97/113)	SD=12.32					SD=0.21		
15. COBALT 2013	-5.60 (1.34)	-3.00 (0.65)						
(213/206)	SD=13.65	SD=6.65						

16. Oriordan-2007			-2.50 (0.85)	-2.80 (1.16)	-3.20 (1.37)			
(155/146)			SD=7.35	SD=9.99	SD=11.81			
17. CREATE 2006	-3.30 (1.27)		-2.19 (0.90)	-3.61 (1.19)				
(142/142)	SD=10.68		SD=7.61	SD=9.99				
18. MIND-IT 2002 a	0.80 (0.85)						1.10 (1.11)	0.00 (0.81)
(86/132)	SD=6.66						SD=8.06	SD=5.88
19. THREAD 2009	-2.16 (1.33)		-2.49 (0.81)				7.56 (3.03)	
(90/96)	SD=9.06		SD=5.49				SD=20.58	
20. CADET 2013		-1.60 (0.63)					3.90 (1.31)	0.20 (1.20)
(275/230)		SD=7.03					SD=14.50	SD=13.53
21. Dowrick 1996	-2.49 (1.33)						5.83 (2.95)	
21. Dowrick 1996	-0.71 (1.39)						6.37 (3.09)	
(139/98;80)	SD=10.01						SD=22.31	
22. Fabre 1992 b			-4.56 (1.90)		-4.30 (1.95)			
22. Fabre 1992 b			-6.07 (1.89)		-6.98 (1.94)			
(36/37;38)			SD=8.11		SD=8.33			
23. Pedersen 2003 a			-1.30 (0.82)		-2.40 (1.21)			
23. Pedersen 2003 a			-2.40 (0.81)		-3.70 (1.20)			
(117/96;99)			SD=5.96		SD=8.78			
24. Demitrack 2001			-3.12 (0.78)		-3.38 (1.11)			
24. Demitrack 2001			-1.14 (0.96)		-2.23 (1.45)			
(68/66;33)			SD=6.35		SD=9.21			
25. Serfaty 2009	-1.91 (1.99)					0.07 (0.07)		
25. Serfaty 2009	-0.03 (1.86)					0.09 (0.06)		
(59/58;55)	SD=10.40					SD=0.34		
26. Titov 2010	-10.86 (2.22)	-5.39 (0.94)						
26. Titov 2010	-11.56 (2.29)	-5.68 (0.96)						
(40/41;46)	SD=10.41	SD=4.33						
27. Richards 2008		-5.02 (1.99)					0.42 (2.86)	1.69 (3.29)
27. Richards 2008		-3.55 (2.05)					1.68 (2.74)	1.79 (3.32)
(27/35;34)		SD=7.58					SD=10.87	SD=11.95
28. Kasper 2006	-4.30 (1.20)		-4.80 (1.11)		-5.30 (1.43)		9.70 (2.55)	3.60 (2.11)
28. Kasper 2006	-4.60 (1.17)		-5.60 (1.07)		-7.00 (1.44)		11.30 (2.36)	1.60 (2.27)
(81/124;119)	SD=8.59		SD=7.20		SD=9.59		SD=19.12	SD=16.48
29. Goldstein 2004			-2.43 (0.85)		-1.94 (1.23)			
29. Goldstein 2004			-3.62 (0.74)		-3.30 (1.09)			
29. Goldstein 2004			-1.23 (0.78)		-1.58 (1.20)			
(88/84;86;84)			SD=7.41		SD=10.77			

30. Dimidjian 2004_1	-1.74 (3.07)	-1.64 (1.61)	
30. Dimidjian 2004_1	-0.79 (2.42)	0.35 (1.78)	
30. Dimidjian 2004_1	0.65 (3.15)	-0.48 (1.87)	
(19/17;28;15)	SD=9.09	SD=5.63	
31. Dimidjian 2004_2	-5.68 (3.31)	-2.96 (2.31)	
31. Dimidjian 2004_2	-8.11 (3.29)	-3.42 (1.97)	
31. Dimidjian 2004_2	-1.50 (4.23)	-3.23 (2.22)	
(22/22;38;21)	SD=11.82	SD=7.26	
32. Loo 2002 a		-2.25 (1.17)	-3.27 (1.42)
32. Loo 2002 a		-2.17 (1.16)	-2.01 (1.48)
32. Loo 2002 a		-0.64 (1.21)	-1.52 (1.48)
32. Loo 2002 a		-2.57 (1.18)	-3.64 (1.41)
(136/144;136;135)		SD=8.44	SD=9.24

A network diagram (Figure 1) shows how many trials report each pair of test outcomes. HAMD17 and BDI were the most common. The proposed method requires that each of the tests included are "connected" to one or more other tests by at least one trial. There were 25 2-outcome studies, 6 3-outcome, and one 5-outcome study. There were 20 studies with 2 arms, 8 with 3 arms, 3 with 4 arms and 1 with 5 arms.

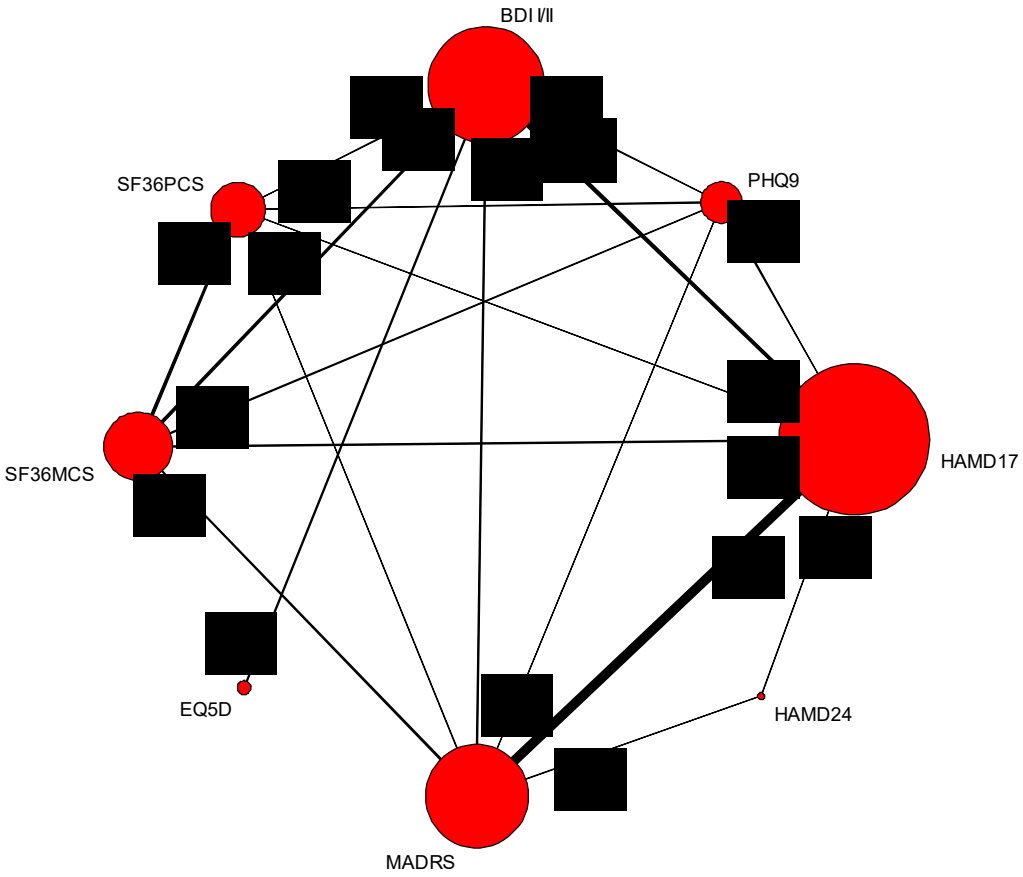

Figure 1: Network of outcomes from the 32 studies considered. Each node represents each outcome with size proportional to the number of trials that include the given outcomes. The thickness of each edge is proportional to the total sample size in each pairwise connection.

The data used in the study were the mean treatment differences after 12 weeks follow-up, or as close to that as possible. Treatment differences and their standard errors were based on mean difference at follow-up if available; otherwise mean differences in change-from-baseline. Similarly, the pooled standard deviation at follow-up was used for standardization if available; otherwise the pooled standard deviation on the changes scores (Table 1).

2.2 Statistical methods

We use statistical models developed previously [15, 16] to allow for simultaneous estimation of both treatment effects on continuous outcomes and "mappings" between treatment effects in a Bayesian framework. The mappings are simply ratios of the underlying treatment effects on their original scales. Here we apply the same model to standardized effects, interpreting the mappings between standardized effects as relative responsiveness ratios.

The model for the treatment effect was a conventional random effects model with study effects δ_{i1} drawn from a normal distribution with mean μ and variance σ^2, whilst *i* denote study. Under the assumption of no difference in efficacy between the different treatments for depression, any variation between treatments would be absorbed into the between-trial variation and (see discussion).

The index 1 on δ_{i1} indicates that this is the standardized treatment effect on the BDI scale; any of the measurement instruments could be selected as outcome 1, but BDI is chosen as it is one of the most frequently used tests. The fact that BDI is the only outcome that can be compared directly to each of the other measures (see Figure 1) also favors its adoption as the reference outcome, as this improves stability and convergence in Markov chain Monte Carlo (MCMC) estimation (see below). Two models for the relationship between the test instruments were examined. In the first it is assumed that the treatment effects on test instruments *h* and *k,* δ_{ih}, δ_{ik} are in the same fixed ratio in every trial (fixed mapping ratio):

$$\frac{\delta_{ik}}{\delta_{ih}} = \beta_{h->k}, \text{ for all } i \qquad (1)$$

In a second model (random mapping ratio) this assumption is relaxed, and ratios are allowed to vary around their mean value:

$$\frac{\delta_{ik}}{\delta_{ih}} = \beta_{i,h->k}, \quad \beta_{i,h->k} \sim N(\beta_{h->k}, \sigma^2_{hk}) \qquad (2)$$

In the random mapping model, we have proposed a constant between-trials coefficient of variation (CV), φ, which is the between-trials standard deviation of mapping coefficients divided by the mean mapping [15, 16], so that $\sigma^2_{hk} = \beta^2_{h->k}\varphi^2$. An important feature of both models, which follows from (1) and (2) is that the ratios must be *transitive*, and *invertible:*

$$\beta_{x->z}=\beta_{x->y}\beta_{y->z}, \text{ and } \beta_{x->z}=\frac{1}{\beta_{z->x}} \quad (3)$$

As a result of these logical constraints, if there are *M* different measurement instruments (in this example *M*=8), there are *M*(*M*-1)/2 (=28) ratios to be estimated, but it is only necessary to estimate *M*-1 (=7) *basic* relative responsiveness parameters [25] , for example the ratios of the first with the other *M-1*, as the remaining 21 parameters are functions of them as shown in (3). The fixed and random relative responsiveness models can be compared in terms of goodness of fit.

Our Bayesian approach places vague prior distributions on the pooled mean treatment effect, μ the between-trial standard deviation of treatment effects, σ^2 the *M*-1 mapping coefficients $\beta_{1->h}$, and, in the random mapping ratio model, the between-trial CV φ. Technical details of the prior distributions are given in Supplementary materials.

Correlations between the effect sizes from the same trial must be taken into account. These can be derived from correlations between original test scores, as described previously [15, 26].Because correlations are not generally reported in trials, we have assumed a correlation matrix based on external information (Table 2). The correlations between the psychiatric measures were based on Handbook of Psychiatric Measures [27] and a recent publication on QoL measures in depression [27]. We examined sensitivity of results to the assumed correlations, by raising or lowering them by a factor of 40%.

Table 2. Assumed within-study correlations between variables. Based on Handbook of Psychiatric Measures [27] and a recent HTA repost on QoL measures in depression [39].

	BDI I/II	PHQ9	HAMD17	HAMD24	MADRS	EQ5D	SF36MCS	SF36PCS
BDI I/II	1	0.60	0.65	0.65	0.65	-0.40	-0.65	-0.05
PHQ9		1	0.60	0.60	0.60	-0.35	-0.60	-0.05
HAMD17			1	0.65	0.65	-0.40	-0.65	-0.05
HAMD24				1	0.65	-0.40	-0.65	-0.05
MADRS					1	-0.40	-0.65	-0.35
EQ5D						1	0.40	0.50
SF36MCS							1	0.10
SF36PCS								1

2.3 Statistical estimation

Estimation was carried out by Markov Chain Monte Carlo (MCMC) using WinBUGs [28] Goodness of fit was assessed via the posterior mean Residual Deviance ($\bar{D}$) as a global goodness of fit statistic [29]. Convergence, based on statistical criteria [30], was achieved within 50,000 iterations. Posterior summaries were based on 100,000 samples over 5 chains with different starting values and after having discarded the first 200,000 samples to arrive at estimates with Monte Carlo error less than 5% of the sample standard deviation for all parameters.

3. RESULTS

Posterior summaries of estimated parameters for the fixed and random relative responsiveness ratios are compared in Table 3. Both fixed and random ratio models fit the data well, with $\bar{D}$ = 112.3 and 107.7 respectively. In a well-fitting model $\bar{D}$ should be approximately equal to the total number of observations, which are 111 in this dataset. The posterior means of the relative responsiveness ratios in the fixed and random ratio are extremely close. The posterior median of the between-study CV of the relative responsiveness ratios is 13%. The CV quantifies the amount of between-study variation of treatment ratios or relative instrument responsiveness and is a direct way of testing whether instrument relative responsiveness varies across studies or can be thought as fixed and equal to 1. The fixed ratio model assumes that instrument relative responsiveness is fixed to 1. In other words under the fixed ratio model, the test instruments are assumed as equal responsive relative to the chosen reference instrument. This assumption is relaxed under the random ratio model. The between-study variation in instrument relative responsiveness is then quantified by the CV parameter. The CV summarizes this variation by expressing the between-study SD of relative responsiveness ratios relative to their mean values and was found to be 13% of their mean values. This is a moderate value, but the 95% Credible Interval (CrI) of 6%-25% is relatively far from zero, indicating statistical grounds for preferring the random ratio model.

We can interpret the relative responsiveness ratios as follows: taking HAMD17 as an example and using the random ratio model, one SD unit treatment effect of BDI is on average equivalent to 1.33 SD unit effect on HAMD17. In other words, HAMD17 is more

responsive to treatment changes than BDI by a factor of 1.33 (95% CrI: 1.04 – 1.67). Evidently, PHQ9 is the most responsive to treatment, and SF-36 PCS the least (Table 3).

Table 3. Posterior summaries of treatment effect, between-trials variation, relative responsiveness ratios, and ranking of relative responsiveness ratios (rank 1 is the most responsive).

	Fixed Ratios	**Random Ratios**	**Ranking estimates: random ratio model**
	Mean (SD)	**Mean (SD)**	***2.5% 50% 97.5%***
Mean treatment effect on BDI (μ)	-0.27 (0.04)	-0.26 (0.04)	
	Median [95% CI]	**Median [95% CI]**	
Between-trial sd (σ)	0.15 [0.10, 0.23]	0.15 [0.09, 0.23]	
Responsiveness relative to BDI			
BDI (reference)	1 -	1 -	[4 6 7]
PHQ9	1.52 [1.17 2.05]	1.53 [1.14 2.14]	[1 1 4]
HAMD17	1.31 [1.04 1.69]	1.31 [1.02 1.72]	[1 3 5]
HAMD24	1.31 [0.80 2.07]	1.30 [0.75 2.22]	[1 3 7]
MADRS	1.29 [1.02 1.69]	1.26 [0.94 1.71]	[1 4 6]
EQ5D	-0.59 [-1.24 -0.12]	-0.58 [-1.35 -0.08]	[3 7 8]
SF36MCS	-1.20 [-1.62 -0.89]	-1.22 [-1.72 -0.87]	[1 4 6]
SF36PCS	-0.36 [-0.92 -0.03]	-0.37 [-0.94 -0.07]	[7 8 8]
Between-trial CV(φ)	-	0.13 [0.06, 0.25]	
Mean Residual Deviance ($\bar{D}$)	112.3	107.7	

The mean (standardized) treatment effect on the BDI scale is -0.26 with posterior standard deviation (0.04). This has no useful interpretation as it is an average of an arbitrary mixture of active treatments against placebo or waitlist. However, in comparison the method allows investigators to report treatment effects on any of the (standardized) scales [16, 31] (A.2). For example the treatment effect on the standardized HAMD17 scale will be -0.26 times 1.33.

The posterior densities of the relative responsiveness ratios relative to BDI can be found in Figure 2. (For ease of interpretation we have ignored the signs of the ratios which simply denote that effective treatment produces an increase on EQ-5D, SF-36 MCS and PCS, and a decrease on PHQ9, BDI, HAMD17, HAMD24, and MADRS).

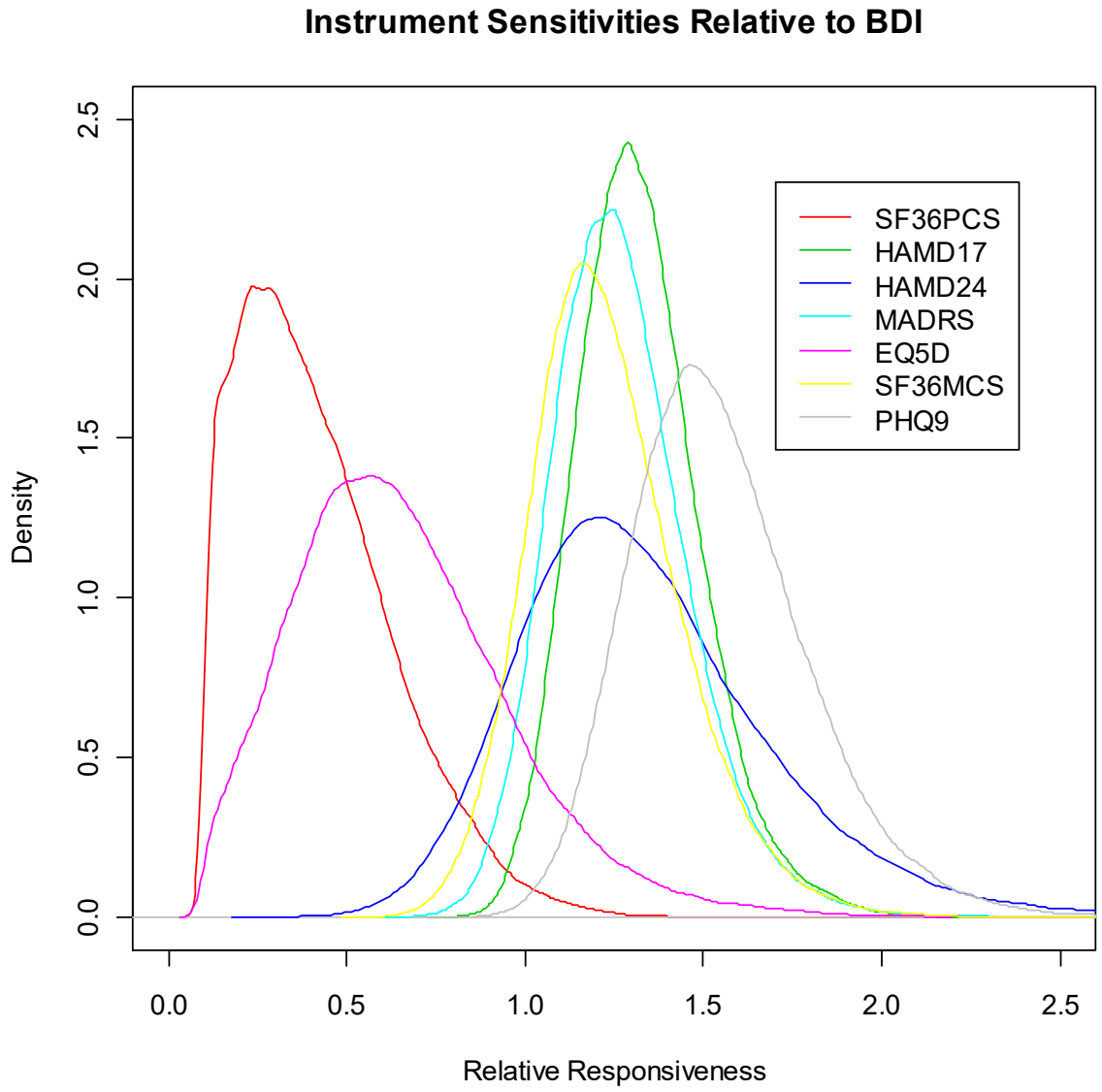

Figure 2: Distribution densities of the relative responsiveness ratios of each test instrument relative to BDI over-laid on the region of common support. (Note: signs have been ignored)

The "rankogram" (Figure 3) and the Box Plot (Figure 4) give a better impression of the reliability of the differences between the instruments in responsiveness. These show that there is considerable uncertainty regarding each instrument's rank. For example PHQ9 is the best (most responsive) but there is still a 25% chance that it is ranked 3rd or lower. All instruments with the exception of EQ5D have very little chance of being less responsive than SF36PCS. The relative responsiveness of the 5 depression-specific measures and SF36MCS is of particular interest. The difference between the cumulative probabilities of the best (PHQ-9) and worst (BDI) is 84%. This can be referred as the distribution of the

expected differences between ranked probabilities drawn from a uniform distribution [32]. The probability of finding a difference this large is less than 5%.

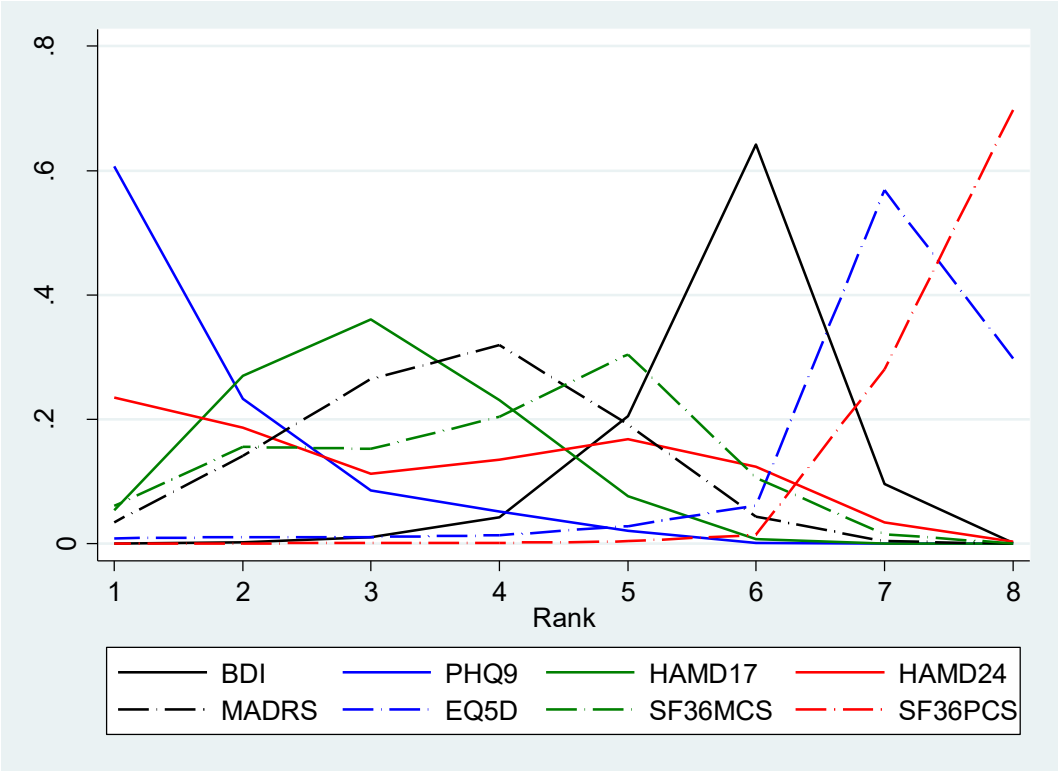

Figure 3: Rankograms of instrument sensitivity to treatment under the random mapping model.

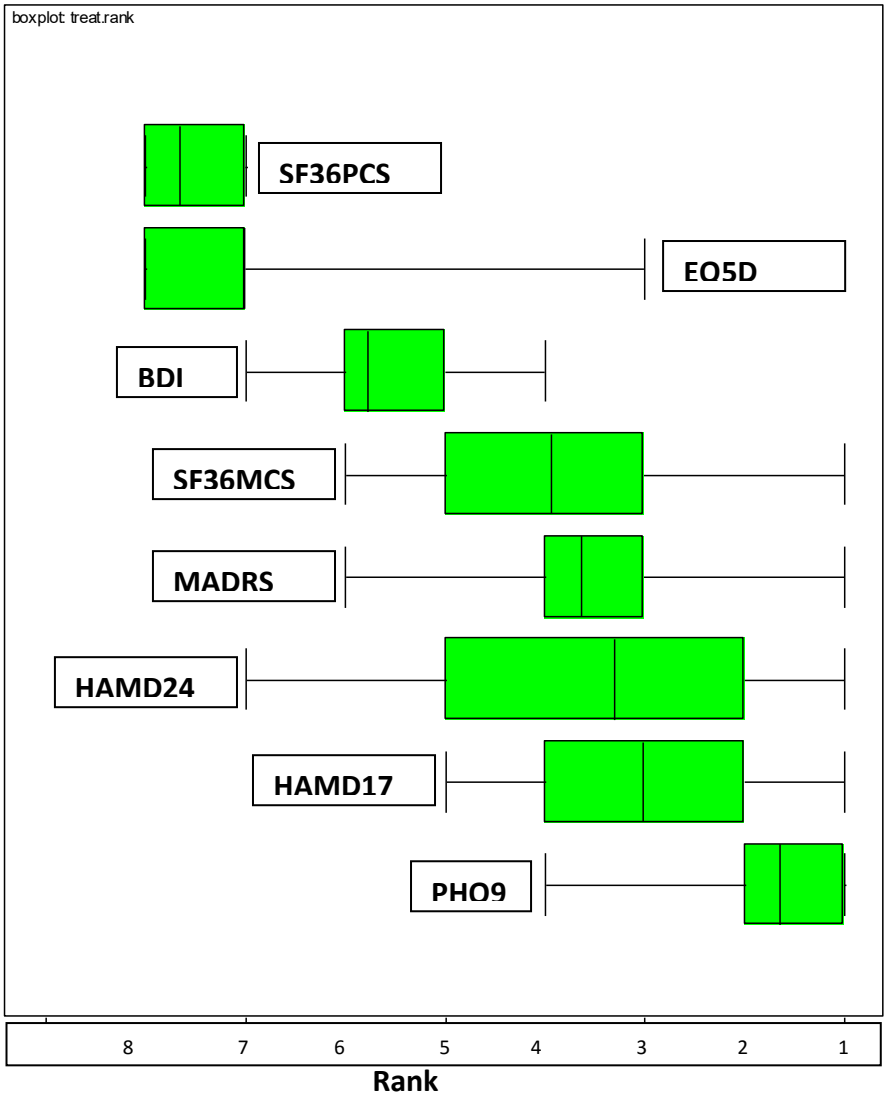

Figure 4: Box-plot comparing instrument specific posterior distribution of rankings.

Sensitivity of results to the correlation between test scores was assessed by raising or lowering correlations by 10% and 40%. In both models this has very little effect on the mean treatment effect, the between trials variation in treatment effects, or the treatment ratios. Only for large increases (40%) in the correlation matrix we noticed an increase to the coefficient of variation of mapping of 6% [i.e. from 13% to 19%] in the random ratios model. However, even this unrealistic variation in the correlations, induced changes in the mean relative responsiveness ratios which ranged between 0.02 – 0.06.

4. Discussion

Comparative instrument responsiveness studies are needed to improve the assessment of mental health interventions and their health economic evaluation as well as inform optimal study design in mental health. This paper demonstrates a new meta-analytic method for

assessing the relative responsiveness of test instruments from a connected network of trials. The method, in essence, carries out an across-trial pooling of the within-trial information on the ratios of standardized effect sizes, interpreting these ratios as measures of relative responsiveness. It takes advantage of the logical requirement that the true ratios must be transitive, meaning that the ratio of test outcomes A and C is the product of the A: B and B: C ratios, and uses this to facilitate an efficient and coherent pooling of the ratio information.

The substantive findings from the study are limited by its small size. However, previous psychometric comparisons of different test instruments for depression have been generally restricted to comparing a very small number of tests, usually two [33-35], based on single studies. Where meta-analysis of several studies has been carried out, this was only when every study included all the tests [13, 14]. Our finding that PHQ-9 was superior to BDI in responsiveness confirms a previous reports making this comparison [35]. However, its high ranking is somewhat surprising as PHQ-9 is a very short questionnaire with only 9 items. According to standard Measurement Theory greater precision and less measurement error should be associated with longer tests with more items [36]. On the other hand, the brevity of the PHQ-9 is considered an advantage for monitoring depression outcomes, as patients' responses may deteriorate in quality if instruments are too long [37].

The finding that generic measures such as EQ-5D and SF-36 PCS are very insensitive to differences caused by treatments for depression is expected from the literature [38, 39]. Relatively high correlations between standard depression test instruments and SF36 mental health have been observed before [40, 41]. A number of studies have showed that SF36-MCS can be a useful screening test for depression based on studies of different sub-populations [42-46].

It is of some interest that there was evidence that relative responsiveness ratios vary from trial to trial though the coefficient of variation was modest, 13%. Similar degrees of between-trial variation have been observed in measures of social anxiety [16] and ankylosing spondylitis [15] . A certain degree of variation is to be expected from the fact that the relationships between the different outcomes are only approximately linear, with some test instruments more sensitive to differences at one end of the continuum, and some at the other [47]. It follows that any heterogeneity in depression severity can be expected to generate random variation in relative responsiveness ratios. Furthermore, variation

between studies in the proportion of total variance attributable to the target construct (depression) is also likely to generate variation in relative responsiveness.

Our method assumes that test instruments have the same responsiveness regardless of treatment. There have been suggestions that the MADRS scale is specifically responsive to the effects of tricyclic amines [13], and that BDI, MDRS and HAMD instruments are differentially responsive to SSRIs and nortriptyline [45]. There are also indications that the HAMD subscales are differentially responsive to SSRIs and SRNIs [14]. Such interactions cannot be ruled out, although a far larger body of data would be required to validate them. For the present we believe it is reasonable to assume that such effects are relatively minor.

Our assumption that differences between measurement instruments in responsiveness are constant across treatments is, of course, far weaker than the assumption routinely made in meta-analyses based on standardized scores from different instruments, namely that all the instruments are ***equally*** responsive. A more general critique of standardization based on these methods was published previously [16].

Previous meta-analytic work has been restricted to bodies of data in which every test instrument has been reported in every study [13, 14], placing a heavy restriction on the number of instruments compared. Murawski and Miederhof [12] assembled a large body of evidence on a range of generic and disease specific instruments. Although they also used standardized effect sizes, they looked at within-trial *differences* between effect sizes, rather than *ratios,* and reported quite extreme levels of between-trial heterogeneity in effect size differences. Relatively few studies have attempted formal statistical inference on differences in comparative responsiveness, and still fewer have taken account of the between-outcome correlation. The method proposed here represents two important technical advances in assessing relative responsiveness: first, by exploiting the logical transitivity relationship between relative responsiveness parameters we can generate a single coherent ranking, based on within-trial information, without requiring every trial to report every outcome. Second, the method explicitly takes account of the within-study correlations between outcomes. Although we were obliged to use external information from the psychometric literature, sensitivity analyses showed that our substantive findings were robust against plausible mis-specification of the correlations.

The proposed method itself has limits. Firstly, because it is based on ratios of effects, it will perform best when applied to trials with large effect sizes. The method could become

unstable if treatment effects were small. Second, it is necessary to include only studies reporting on two or more outcomes, forming a connected network of outcomes. Single outcome studies could be included only by introducing cross-study comparisons, which is likely to increase the between-study variance of ratios very considerably [12].

A further limitation is that, while the choice of reference outcome does not alter the fixed effects statistical model, this is not strictly true in the case of the random ratios model. If, for example, HAMD-17 is chosen as the reference standard, then the accuracy of the relative responsiveness ratios can be affected. While this is not a desirable property of the model, relative responsiveness point estimates with HAMD-17 as the reference outcome differ by no more than a factor of 1.03. Similar insensitivity of results to choice of reference test has been reported previously [16]. However, we are working on alternative parameterizations which avoids this technical short-coming.

In spite of these limitations we believe these proposals represent a methodological advance in the study of relative responsiveness.

The current study focuses on relative responsiveness to treatment, ascertained in randomized trials. These have the advantage of being based on causal changes due to the treatments, while in non-trial settings it less easy to rule out the influence of other factors on change scores that are not related to treatment. At the same time, the trial setting defines what it is that "responsiveness" is supposed to measure [1], namely the effect of the treatment. However, there is no reason why similar methods could not be applied to suitably standardize longitudinal measures of responsiveness, as long as the correlation structures between different tests at the same time point, between the same test at different time points, and between different tests at different time points, are carefully taken into account. Whether or not relative responsiveness in these two contexts is the same can only be assessed from very large ensembles of data.

We noted above the connection between relative responsiveness ratios and the "mappings" between un-standardized treatment effects on different test instruments [10]. Extensive use of mapping from disease specific to generic measures such as EQ-5D is a routine feature of health technology assessment, undertaken to estimate health gains on a monetized scale [42-44].

There is a large literature on mapping between test instruments, also referred to as cross-walking, test-linking, or test equating [48, 49] mainly oriented to educational measurement. These methods could, again, be readily applied to standardized data. Among the methods available Item Response Theory, or Rasch Analysis [50], is a particularly powerful technique that can be used to highlight differences in the relative responsiveness of test instruments in particular parts of their scales. However, all these methods require individual patient data on each of the subscales from which the test instruments are formed. This is seldom available for more than a single dataset, and our results here show that it is important to account for between-study variation.

For this reason, we feel that the use of aggregate trial data to estimate relative responsiveness, is a simple and practical method, although one that could be developed further.

Acknowledgements

This paper is independent research funded by the National Institute for Health Research (Programme Grants for Applied Research, What are the indications for Prescribing ANtiDepressAnts that will lead to a clinical benefit: PANDA, RP-PG-0610-10048). The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

KSB was funded by the National Institute for Health Research School for Primary Care Research (NIHR SPCR). The NIHR SPCR is a partnership between the Universities of Birmingham, Bristol, Keele, Manchester, Nottingham, Oxford, Southampton and University College London.

References

[1] Terwee C, Dekker F, Wiersinga W, Prummel M, Bossuyt P. On assessing responsiveness of health-related quality of life instruments: guidelines for instrument evaluation. Quality of Life Research. 2003;12:349-62.
[2] Beaton DE, Bonmbardier C, Katz JN, Wright JG. A taxonomy for responsiveness. Journal of Clinical Epidemiology. 2001;54:1204-17.
[3] Schuck P. Designs und Kennziffern zur Ermittlung der Änderungssensitivität von Fragebogen in der gesundheitsbezogenen Lebensqualitätsforschung ([Designs and statistics for assessing responsiveness of questionnaires in health-related quality of life research]) . [Designs and statistics for assessing responsiveness of questionnaires in health-related quality of life research]. Z Med Psychol. 2000;9:125-30.
[4] Hevey D, McGee HM. The effect size statistic: useful in health outcomes research ? Journal of health psychology. 1998;3:163-70.

[5] Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation. Control Clin Trials. 1991;12:142S-58S.
[6] Fitzpatrick R, Davey C, Buxton MJ, DR J. Evaluating patient-based outcome measures for use in clinical trials. Health Technology Assessment1998.
[7] Cohen J. Statistical power analysis for the behavioral sciences. New York: Academic Press; 1969.
[8] Hedges L. Distribution theory for Glass's estimator of effect size and related estimators. Journal of Educational Statistics. 1981;6:107-28.
[9] Borenstein MA, Hedges L, Higgins J. Introduction to Meta-analysis. Chichester: Wiley; 2009.
[10] Fayers PM, Machin D. Quality of life. The assessment, analysis and interpretation of patient-reported outcomes. The Atrium, Southern gate, Chichester, West Sussex PO19 8SQ, England: John Wiley % Sons Ltd; 2007.
[11] Wiebe S, Guyatt G, Weaver B, Matijevic S, Sidwell C. Comparative responsiveness of generic and specific quality-of-life instruments. Journal of Clinical Epidemiology. 2003;56:52-60.
[12] Murawski MM, Miederhoff PA. On the generalizability of statistical expressions of health related quality of life instrument responsiveness. Quality of Life Research. 2007;7:11-22.
[13] Faries DE, Herrera J, Ravamajhi J, DeBrota D, Demitrack M, Potter WZ. The responsiveness of the Hamilton Depression rating scale. J Psychiatric Res. 2000;34:3-10.
[14] March JS, Entusah AR, Rynn M, Albano AM, Tourian KA. A randomized controlled trial of venlafaxine ER versus placebo in pediatric social anxiety disorder. Biological Psychiatry. 2007;62:1149-54.
[15] Lu A, Kounali D, Ades AE. Simultaneous multi-outcome synthesis and mapping of treatment effects to a common scale. Value Health. 2014;17:280-7.
[16] Ades AE, Lu G, Dias S, Mayo-Wilson E, Kounali D. Simultaneous synthesis of treatment effects and mapping to a common scale: an alternative to standardisation. Research Synthesis Methods. 2015;6:96-107.
[17] Ades AE, Lu G, Madan JJ. Which health-related quality-of-life outcome when planning randomised trials: disease-specific or generic, or both? A common factor model. Value in Health. 2013;16:185-94.
[18] Lu G, Brazier JE, Ades AE. Mapping from disease-specific to generic health-related quality-of-life scales: a common factor model. Value in Health. 2013;16:177-84.
[19] Beck AT, Steer RA, Carbin MG. Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation. Clinical Psychology Review. 1988;8:77-100.
[20] Spitzer RL, Kroenke K, Williams JB. Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. Primary Care Evaluation of Mental Disorders. Patient Health Questionnaire. Jama. 1999;282:1737-44.
[21] Hamilton M. A rating scale for depression. J Neurol Neurosurg Psychiatry. 1960;23:56-62.
[22] Jaeschke R, Singer J, Guyatt G. Measurement of health status: Ascertaining the minimal clinically important difference. Controlled Clinical Trials. 1989;10:407-15.
[23] Szende A, Williams A. Measuring self-reposrted Population Health: An international perspective based on EQ-5D. EuroQol Group; 2004.
[24] Ngo-Metzger Q, Sorkin DH, Mangione CM, Gandek B, Hays RD. Evaluating the SF-36 Health Survey (Version 2) in Older Vietnamese Americans. Journal of aging and health. 2008;20:420-36.
[25] Eddy DM, Hasselblad V, Shachter R. An introduction to a Bayesian method for meta-analysis: the confidence profile method. Medical Decision Making. 1990;10:15-23.
[26] Wei Y, Higgins J. Estimating within-study covariances in multivariate meta-analysis with multiple outcomes. Statistics in Medicine. 2013a;32:1191-205.
[27] Rush AJ. Handbook of psychiatric measures. 2nd ed. ed. Washington, DC: American Psychiatric Pub.; 2008.
[28] Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS -- a Bayesian modelling framework: concepts, structure, and extensibility. Statistics and Computing. 2000;10:325-37.
[29] Spiegelhalter D, Best N, Carlin B, van der Linde A. A Bayesian measure of model complexity and fit. Journal of the Royal Statistical Society (B). 2002;64:583-616.

[30] Brooks SP, Gelman A. Alternative methods for monitoring convergence of iterative simulations. Journal of Computational and Graphical Statistics. 1998;7:434-55.
[31] Lu G, Kounali D, Ades AE. Simultaneous multi-outcome synthesis and mapping of treatment effects to a common scale. Value in Health. 2014;17:280-7.
[32] Gentle JE. Computational Statistics: Springer; 2009.
[33] Cameron IM, Crawford JR, Lawton K, Reid IC. Psychometric comparison of PHQ-9 and HADS for measuring depression severity in primary care. British Journal of General Practice. 2008;58:32-6.
[34] Adler MW, KHetta J, Isacsson G, Brodin U. An item-response-theory evaluation of three depression assessment instruments in a clinical sample. BMC Medical Research Methodology. 2012;12.
[35] Titov N, Dear B, McMillan D, Anderson T, Zou J, Sunderland M. Psychometric comparison of the PHQ-9 and BDI-II for measuring response during treatment of depression. Cogn Behav Ther. 2011;40:126-36.
[36] Spearman CC. Correlation calculated from faulty data. British Journal of Psychology. 1910;3:271-95.
[37] Löwe B, Unützer J, Callahan CM, Perkins AJ, Kroenke K. Monitoring Depression Treatment Outcomes With the Patient Health Questionnaire-9. Medical Care. 2004;42:1194-201.
[38] Peasgood T, Brazier J, Papaioannou D. A systematic review of the validity and responsiveness of EQ-5D and SF-6D for depression and anxiety. HEDS Discussion paper 12/15. Unpublished. 2012.
[39] Brazier J, Connell J, Papaioannou D, Mukuria C, Mulhern B, Peasgood T, et al. A systematic review, psychometric analysis and qualitative assessment of generic preference-based measures of health in mental health populations and the estimation of mapping functions from widely used specific measures. Health Technol Assess. 2014;18.
[40] Kristjánsdóttir J, Olsson GI, Sundelin C, Naessen T. Could SF-36 be used as a screening instrument for depression in a Swedish youth population? Scandinavian Journal of Caring Sciences. 2011;25:262-8.
[41] Elliott TE, Renier CM, Palcher JA. Chronic Pain, Depression, and Quality of Life: Correlations and Predictive Value of the SF-36. Pain Medicine. 2003;4:331-9.
[42] Longworth L, Rowen D. NICE DSU Technical Support Document 10: The use of mapping methods to estimate health state utility values, avaiulable at http://www.nicedsu.org.uk. 2011. p. 1-31.
[43] Brazier JE, Yang Y, Tsuchiya A, Rowen DL. A review of studies mapping (or cross-walking) non-preference based measures of health to generic preference-based measures. European Journal of Health Economics. 2010;11:215-25.
[44] Longworth L, Rowen D. Mapping to obtain EQ-5D utility values for use in NICE health technology assessments. Value in Health. 2013;16:202-10.
[45] Uher R, Maier W, Hauser J, Marušič A, Schmael C, Mors O, et al. Differential efficacy of escitalopram and nortriptyline on dimensional measures of depression. The British journal of psychiatry : the journal of mental science. 2009;194:252-9.
[46] Lincoln R. The SF36 Health Survey: A summary of Responsiveness to clinical interventions. QualityMetric, Inc.; 2000.
[47] Olino TM, Yu L, Klein DN, Rohde P, Seeley JR, Pilkonis PA, et al. Measuring depression using item response theory: an examination of three measures of depressive symptomatology. International Journal of Methods in Psychiatric Research. 2012;21:76-85.
[48] Dorans NJ, Pommerich M, Holland PW. Linking and aligning scores and scales. New York: Springer; 2007.
[49] Kolen MJ, Brennan RL. Test equating, scaling and linking: methods and preactices. New York: Springer; 1994.
[50] Streiner D, Momran G. Health measurment scales: A practical guide to their development and use. New York: Oxford University Press; 2003.